REVIEW ARTICLE

Reproducibility of preclinical data: one man's poison is another man's meat

Anton Bespalov[1,2*], Christoph H. Emmerich[1], Björn Gerlach[1] and Martin C. Michel[1,3]

[1] Partnership for Assessment and Accreditation of Scientific Practice, Am Aukopf 14/1, D-69118 Heidelberg, Germany
[2] Valdman Institute of Pharmacology, Pavlov Medical University, ul. Lva Tolstogo, 6-8, 197022 St. Petersburg, Russia
[3] Department of Pharmacology, Johannes Gutenberg University, Saarstraße 21, D-55122 Mainz, Germany

Abstract: Limited reproducibility of preclinical data is increasingly discussed in the literature. Failure of drug development programs due to lack of clinical efficacy is also of growing concern. The two phenomena may share an important root cause — a lack of robustness in preclinical research. Such a lack of robustness can be a relevant cause of failure in translating preclinical findings into clinical efficacy and hence attrition, and exaggerated cost in drug development. Apart from the study design and data analysis factors (e.g., insufficient sample sizes, failure to implement blinding, and randomization), heterogeneity among experimental models (e.g., animal strains) and the conditions of the study used between different laboratories is a major contributor to the lacking of robustness of research findings. The flipside of this coin is that the understanding of the causes of heterogeneity across experimental models may lead to the identification of relevant factors for defining the responder populations. Thus, this heterogeneity within preclinical findings could be an asset, rather than an obstacle, for precision medicine. To enable this paradigm shift, several steps need to be taken to identify conditions under which drugs do not work. An improved granularity in the reporting of preclinical studies is central among them (i.e., details about the study design, experimental conditions, quality of tools and reagents, validation of assay conditions, etc.). These actions need to be discussed jointly by the research communities interested in preclinical data robustness and precision medicine. Thus, we propose that a lack of robustness due to the heterogeneity across models and conditions of the study is not necessarily a liability for biomedical research but can be transformed into an asset of precision medicine.

Keywords: animal models, data reproducibility, heterogeneity, precision medicine, translational research

*Correspondence to: Anton Bespalov, PAASP GmbH, Am Aukopf 14/1, D-69118 Heidelberg, Germany; Email: anton.bespalov@paasp.net

Received: April 6, 2016; **Accepted:** June 18, 2016; **Published Online:** September 23, 2016

Citation: Bespalov A, Emmerich C H, Gerlach B, *et al*., 2016, Reproducibility of preclinical data: one man's poison is another man's meat. *Advances in Precision Medicine*, vol.1(2): 1–10. http://dx.doi.org/10.18063/APM.2016.02.001.

Introduction

Private and public investment in fundamental biomedical research continues to be strong; concomitantly, biomedical researchers now have more sophisticated tools at hand than ever before. This combination should offer unparalleled opportunities for the discovery and development of new and potentially transformative therapeutics. However, the overall productivity of drug research and development, i.e., the number of new therapeutics developed per dollar invested, has nonetheless significantly declined over the past several decades[1]. One driver of low productivity is the increasing cost of drug development that is itself driven, for example, by increasing demands from regulatory authorities and health technology assessment bodies. A possibly even bigger cost driver is (late stage) attrition[2]. Attrition rates in

the clinical phase remain high[3], and lack of efficacy has become the most important reason for attrition[4]. There may be a number of reasons for the failure to translate promising preclinical data into clinical efficacy[5,6]. Animal models of disease are the cornerstones of drug development[7] but in many cases have been insufficiently validated for being predictive of efficacy in patients[8,9]. Moreover, potential therapeutic targets, which are in part derived from such models, may have been insufficiently validated[10].

Limited validation of both animal models and proposed drug targets appears to be at least partly related to a lack of reproducibility of preclinical data[11,12]. Accordingly, this lack of reproducibility has become a major concern for funding agencies such as the National Institute of Health (NIH)[13]. For drug discovery and development, failure to reproduce research findings may result in longer drug research and development times, thereby leading to increased costs[14], abandoning of research programs and, as a financial and ethical worst case, to lack of efficacy in clinical proof-of-concept studies. Concerns about the reproducibility of preclinical data have triggered discussions around various aspects of good research practice such as transparent data analysis[15,16], reporting on the use of laboratory animals[17], and reassessment of data publication guidelines by the peer-reviewed literature venues[18].

While the lack of reproducibility is not the only factor leading to poor prediction of clinical efficacy based on preclinical data[19], addressing it is critical for restoring the self-correcting nature of science[20]. In this review, we argue that certain cases of what is called as "lack of data reproducibility" could be converted into exciting discoveries leading to innovative personalized medicine.

Data Reproducibility vs Robustness

Given that the term "reproducibility" itself has caused quite a lot of confusion, it therefore needs to be defined. It has been suggested that the follow-up experiments for an initial set of data fall into one of two categories — replication or robustness tests[21] (Table 1). Replication follow-up tests are conducted using the same methods and population as the original study, and usually meant to evaluate measurement error. Replication tests use the same sampling distribution for parameter estimates; therefore, reduce the conditions for discrepancy result from random chance, error, or inappropriate handling of data (Table 1). In contrast, robustness tests are conducted using different methods or on a new sample drawn from a different population (Table 1) and evaluate the generalizability of research findings. Evaluation of the generalizability of the phenomenon allows a result to be assessed based on differences between testers under different conditions and sampling distributions. In preclinical drug research and development, discussions of "reproducibility" have focused nearly exclusively on tests of robustness. Therefore, in the subsequent discussion, we refer to the robustness of preclinical data rather than "reproducibility".

Figure 1. A proposed definition to distinguish replication and robustness tests.

	Sampling distribution for parameter estimates	Sufficient conditions for discrepancy	Types	Methods in follow-up study versus methods *reported* in original: Same specification	Same population	Same sample	Examples
Replication	*Same*	*Random chance, error, or fraud*	Verification	*Yes*	*Yes*	*Yes*	*Fix faulty measurement, code, dataset*
			Reproduction	*Yes*	*Yes*	*No*	*Remedy sampling error, low power*
Robustness	*Different*	*Sampling distribution has changed*	Reanalysis	*No*	*Yes*	*Yes/No*	*Alter specification, recode variables*
			Extension	*Yes*	*No*	*No*	*Alter place or time; drop outliers*

The "same" specification, population, or sample means the same as *reported* in the original paper, not necessarily what was contained in the code and data used by the original paper. Thus for example if code used in the original paper contains an error such that it does not run exactly the regressions that the original paper said it does, new code that fixes the error is nevertheless using the "same" specifications (as described in the paper).

*Reproduced from[21] with permission.

Data Robustness from the Generalizability Perspective

Most research groups in academia and industry have their preferred protocols, suppliers of tools, reagents and cell lines or animals and, in clinical research, preferred sources and types of study subjects. These factors are dictated by a number of budgetary, logistic, historical, and other practical considerations. They help to standardize methods and test conditions within a laboratory. Thereby, they have a positive impact on the "local" (within-laboratory) probability of success of a research project, as reflected by the outcome of statistical analyses (i.e., standardization efforts are aimed at achieving "statistical generalizability"[22]). However, they may reduce the probability that other investigators, applying minor variations of a protocol, and performing the same experiments at a different location with that location's preferred standards, and obtain the same result. While multi-center trials applying a standardized protocol have become the benchmark in clinical research, their application to preclinical research is only slowly emerging[23].

This uniqueness of each research environment makes the comparisons between results generated across laboratories very important. Seen from the perspective of the development of drugs for typically highly heterogeneous patient populations, it is probable that data generated by one laboratory is more likely to be successfully translated when similar findings are also obtained under the disparate conditions of other laboratories; thus, establishing the generalizability of the results ("scientific generalizability"[22]).

The sensitivity of preclinical assays for a given pharmacological target often depends on assay conditions. This could bias a project work towards having positive results if careful benchmarking is not performed. Therefore, it is essential that methods for defining data generalizability are put into practice. The guiding assumption for this is that the broader the range of circumstances and laboratory environments in which preclinical efficacy can be demonstrated, the higher the likelihood of detecting efficacy signals in the clinical studies that include a broad spectrum of patients. For instance, the clinically successful analgesic morphine anxiolytic diazepam works reliably in both males and females of various strains and species of laboratory animals under most, if not all, laboratory conditions. The question is whether there is value in data that can be generated only under certain conditions (e.g., in one laboratory but not in others). In this report, we argue that, in some cases, the apparent limited robustness of research findings may open the road in identifying a novel approach to personalized medicine.

Generalizability and Precision Medicine

The concept of personalized medicine is based on the heterogeneity of response to drug treatment, i.e., a drug may not work in all patients presented with a given condition. For instance, both muscarinic acetylcholine receptor antagonists and β_3-adrenoceptor agonists have been shown to be effective treatments for the overactive bladder syndrome; however, each of these two drug classes has a limited responder rate, and these responder populations overlap only partly[24,25]. It follows that a drug effective in a subgroup of patients with a given disease may therefore be just partially effective or even ineffective when being tested in a broader population. Under such circumstances, a study may be declared as "negative," adding to the mounting evidence of preclinical-to-clinical translation failures.

Since the completion of the Human Genome Project, major advances in genome technology have led to an exponential decrease in sequencing costs[1]. In some areas of medicine, patients have benefited from the development of new drugs with labels that now include pharmacogenomic information. Patients with melanoma, leukemia, or metastatic lung, breast, or brain cancers are now routinely offered a "molecular diagnosis" (e.g., BRAF-positive melanoma; EGFR-positive or ALK-positive non-small cell lung cancer), and therefore, may receive tailored treatments that can greatly improve the chances of survival.

Progress made in the cancer field is based on significant investments into research on disease biology. However, there are areas of medicine, in which the only available alternative treatment option is when, if no clinical response is observed to one drug, patients are switched to another, and so on, following a conventional trial-and-error approach. This is especially true for the field of psychiatry, where diagnostic categories are not based on biological mechanisms and no tests are available to support the diagnoses. As the trial-and-error approach will certainly not work for the development of novel drugs because clinical studies are too expensive and follow stringent ethical aspects,

one should look for other sources of information which may instruct the development of effective patient-centric treatments. Evidence on drug treatment response heterogeneity in preclinical studies may be one such example of information that is worth being collected and carefully analyzed.

Limited Generalizability of Preclinical Findings: Genetic Factors

The strain or even sub-strain of laboratory animals is known to affect the development and expression of the various phenomena that are the subject of investigation in preclinical research. For instance, the inbred spontaneously hypertensive rat is the most frequently used animal model in arterial hypertension research[26], but animals from different suppliers exhibit genetic and phenotypic heterogeneity[27]. Models of disease-like processes and conditions are frequently reported to require a certain strain of animals. In some cases, differences between strains may have an obvious physiological explanation (e.g., albino rats have compromised vision and may not be used in tasks dependent upon intact vision[28]). In other cases, differences are likely to be more complex, require deeper evaluation and may be very instrumental in developing valid disease models to study novel treatment approaches. For example, Bottger and colleagues[29] have compared the effects of hypercholesterolemic diets in 12 inbred strains of rats and identified hyper-, normo- and hypo-responders. Similarly, the effects of ovariectomy on bone loss varies among strains of mice and this was argued to be relevant for the development of models for postmenopausal osteoporosis and preclinical testing of potential therapies[30,31].

Since the disease models are often designed using a specific and often inbred strain of animal, the effects of drugs are much less likely to be compared between different strains. Nevertheless, the available examples suggest that drug effects may be subject to strain-dependence. For example, the impact of the immunosuppressants cyclosporin A and tacrolimus on retinal ganglion cell survival and axonal regeneration appear to be more likely in Fischer F344 rats than in Lewis rats[32]. Given that it is the Lewis rats, and not the F344 rats, that are vulnerable to inflammatory aspects of autoimmune insults, the neuroprotective and regenerative effects of agents such as cyclosporin A and tacrolimus may be of particular relevance to specific subpopulations of patients with autoimmune disease (e.g., progressive forms of multiple sclerosis).

Evaluating the strain-dependence of the effects of established and investigational treatments may also deliver potentially useful information. For example, the glucocorticoid dexamethasone, the histamine H_1 receptor antagonist pyrilamine, and the novel histamine H_4 receptor antagonist JNJ7777120 have been compared in a mouse model of acute skin inflammation induced by local application of croton oil[33]. While dexamethasone and pyrilamine induced significant anti-inflammatory effects in CD-1 mice and NMRI mice, JNJ777720 was effective only in CD-1 mice[33], a finding likely to affect the translateonal potential of these preclinical data in clinical settings.

Drug safety is just as important as therapeutic efficacy for precision medicine. Drug safety defines not only the dose range to be studied in humans but also the subpopulations of patients who may or may not be susceptible to drug's adverse effects. However, strain-dependent differences in a drug's toxicological profile are not addressed routinely in conventional preclinical safety evaluation programs. Certain findings that preclude further development of a drug candidate may turn out to be strain-specific and of limited or absence of clinical relevance. For example, acrylamide is a commonly used industrial chemical which was suggested to have carcinogenic potential based on studies conducted exclusively in the Fischer F344 rat, but these predictions have since been challenged by studies using other rat strains[34].

Although the strain of research subject may be responsible for the observation of the differences in the study outcome, what is far less frequently emphasized is that factors such as sub-strains and breeding centers supplying research animals could also play a role in the outcome of the study[27]. For example, in the model of dexamethasone-induced osteonecrosis, aside from the known strain-specific susceptibility dexamethasone treatment, there are also sub-strain-specific differences in the adverse effects, i.e., stronger effects in BALB/cJ compared with BALB/cAnNHsd[35]. If the name of the (sub)strain is not explicitly stated in the published report, inconsistencies in results between studies may sometimes be attributed to different suppliers of animals. For instance, Palm and colleagues[36] reported differences in basal and ethanol-induced levels of opioid peptides in Wistar rats from five different suppliers, but this paper does not identify the used strains according to suggested nomenclature[37]. For Wistar rats in particular, this information would have been of great importance given

the sub-strains that were developed and maintained by different breeders (Figure 2), which may have unique behavioral and biochemical characteristics. In the presentation at the 2006 meeting of the Society for Neuroscience, Lindemann and colleagues[38] demonstrated that one of the common Wistar rat sub-strains has a significantly reduced expression of metabotropic glutamate receptor 2 ($mGlu_2$) and, therefore, is less likely to respond to $mGlu_{2/3}$ receptor agonist treatment. Sub-strains of Wistar rats were also shown to exhibit heterogeneous expression of $mGlu_2$ receptors and the Wistar sub-strains with reduced $mGlu_2$ receptor expression also exhibited an anxiety phenotype[39]. As several compounds targeting this receptor system are undergoing clinical development, analysis of the heterogeneity of $mGlu_2$ receptor expression among these rat sub-strains may be important to predict drug treatment response. Accordingly, the Animal Research: Reporting of *In Vivo* Experiments (ARRIVE) guidelines now call for the reporting of animal species, strain, gender, and supplier[17].

Limited Generalizability of Preclinical Findings: Environmental Factors

Laboratory and animal housing environments are arguably the richest source of factors that may appear difficult to control; often neglected but are likely to have an impact on the generalizability of results obtained in preclinical studies. Similar to what was discussed above for genetic factors, analysis of environmental influences may also guide the development of personalized medicine.

Environment is traditionally viewed as an important aspect of model development and treatment efficacy in psychiatry. Indeed, there is a large body of evidence arguing that environmental influences have a significant impact on the performance of animal models in psychiatry (e.g., ventral hippocampus lesion model in rats[40]). Clinical research points at the combined operation of genetic and environmental factors in the development of complex disease states such as major depressive disorder[41]. It is expected that these interactions delineate specific biological subtypes of depression and that individuals with such pathophysiological distinct types of depression will likely respond to different treatments[41].

In preclinical neuroscience, environmental factors are not limited to handling and housing conditions which may be more or less stressful for the study animals and therefore, associated with the outcome of the study. Equally, if not more important, are infections common in animal houses. Health status is rarely reported in scientific publications despite the fact that even a subclinical infection may have a major impact

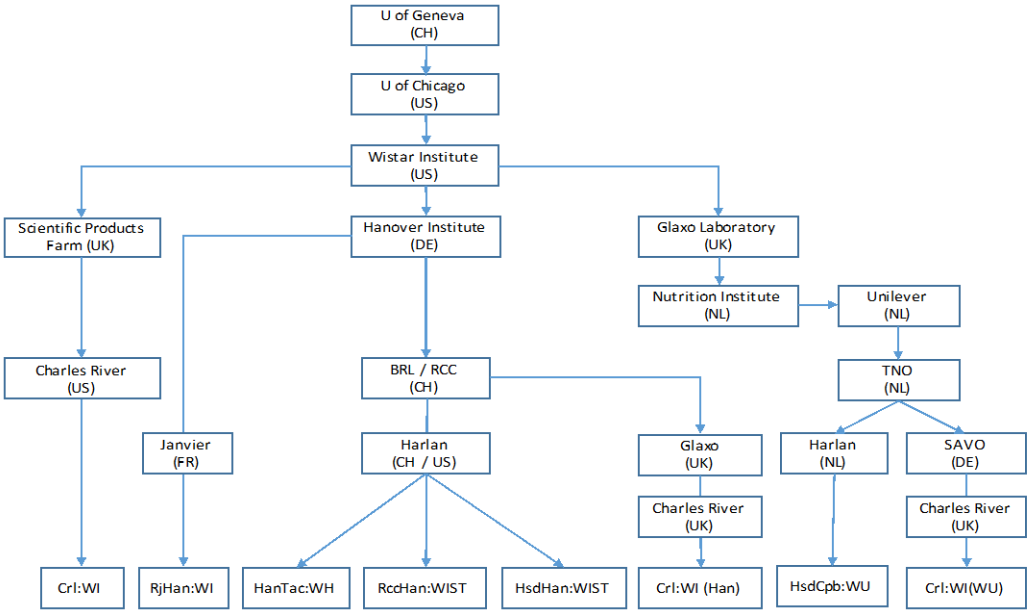

Figure 2. The origins and relationships of Wistar rats available from several commercial sources. Conventional name of the strain is shown in the bottom row (following the information provided by the respective breeders).

on the outcome of preclinical studies in various fields. For example, in the field of neurodegenerative disorders, planned and controlled experimental manipulation (e.g., intracerebral injection of a lesion-inducing agent) may run unintentionally and unknowingly concurrently with subclinical encephalitis due to common mouse virus infections (e.g., mouse hepatitis virus or Theiler's encephalomyelitis virus). Investigational treatment may work well under those conditions but has no impact under "cleaner" conditions. An analogy of this in *in vitro* research is the lack of reporting of potential mycoplasma contamination of cell lines under investigation. If the reasons for differing responses are not known, such observations will be yet another example of the unreliability of preclinical science. If the reasons were known and understood, this could lead to a specific "tailored" use of the studied drug.

Limited Generalizability of Preclinical Findings: The Reverse Translation Approach

There is an overall strong bias towards publishing positive data. Negative data, which might be important in the context of the current discussion, are rarely published and there are three main reasons for this. First, there is no established process to encourage disclosure of such data. Second, such data may often simply not exist because there are many factors that limit motivation to repeat previously published studies — e.g., bioethical considerations or associated costs. Third, authors feel that specifically high profile journals are unlikely to accept negative data.

Therefore, there is a need to stimulate research that would be aimed at establishing heterogeneity of specific phenomena in preclinical research. Clinical evidence of a poor or insufficient response to investigational treatment may serve this purpose (especially, in those cases where obvious reasons for failing preclinical-to-clinical translation can be excluded). For example, κ-opioid receptor antagonists were seen as potential antidepressant drugs, but clinical trial results are so far rather disappointing[42] and may prompt re-evaluation of preclinical evidence. In this vein, κ-opioid receptor antagonists produced antidepressant-like effects in the Wistar-Kyoto but not Sprague-Dawley rats, while a reference agent, the clinically used antidepressant desipramine, was effective in both strains[43]. Deeper analysis of such strain-specific effects is warranted by unsatisfactory clinical efficacy and may lead to the establishment of markers identifying responder population (in both animals and humans).

Another example of a novel drug, in which the results of clinical testing were disappointing, is the $mGlu_{2/3}$ receptor agonist pomaglumetad, which was developed to treat schizophrenia. The Phase III program has been halted in view of insufficient clinical efficacy, which was due, at least in part, to the inability to identify a treatment responder population[44]. As mentioned above, there are preclinical reports on rat sub-strains with reduced expression of $mGlu_2$ receptor and reduced responses to $mGlu_{2/3}$ receptor agonist treatment[38,39]. Furthermore, it has been shown that animals with low levels of $mGlu_2$ receptor expression in the brain are characterized by a higher propensity to drink alcohol[45], and reduced levels of $mGlu_2$ receptor expression have also been noted post mortem in the brains of the human subjects with alcohol dependence[46]. These findings need to be confirmed but it may well be that $mGlu_2$ receptor expression varies in both animals and humans, and this may be associated with a differential response to treatment with $mGlu_2$ receptor agonists and positive allosteric modulators and, therefore, could be used to establish companion diagnostics to enable the personalized use of this class of drugs.

An example of successful reverse translation is the use of β_3-adrenoceptor agonists in the treatment of obesity and type 2 diabetes[47]. Based largely on findings in rodents, several pharmaceutical companies embarked on drug development programs for β_3-adrenoceptor agonists in the 1980s. However, all of these programs failed. It later became clear that the key reason for failure was a difference in the presence of target tissue (brown adipose fat) and of target function (β_3- *vs.* β_1- and β_2-adrenoceptor role in lipolysis) between rodents and humans[47]. Moreover, this is also an example for the role of publication bias, as only one of the various failed β_3-adrenoceptor agonist programs in obesity and type 2 diabetes has published its negative proof-of-concept data[48].

Limited Generalizability of Preclinical Findings: Path Forward

In the clinic, the therapeutic response to medication is known to vary between patients; this is the case for established drugs and may be a contributing factor to the many failures in clinical studies of novel investigational agents that have been tested without any patient stratification strategy. In contrast, preclinical reports on heterogeneity in treatment response are rather

rare. Understanding the reasons for such a lack of information also suggests the steps that are to be taken to generate missing evidence.

First, if both positive and negative data are generated in the same lab using, for example, different strains of animals, a report may indeed be published arguing for strain-dependent effects of an experimental manipulation (e.g., a drug administration). However, if a laboratory generates negative data only, such results may stay unpublished, preventing the scientific community from access to this information. In an extreme case, a laboratory may even report only the positive findings on the animal strain yielding these positive results only, not disclosing those on the strain yielding the negative results. Thus, an obvious step to be taken is to establish a mechanism whereby negative data become shared. Several examples of such mechanisms include publication portals dedicated to negative data (e.g., F1000's preclinical reproducibility channel[49]), pre-print archives (e.g., bioRxiv[50]), online forums for scientific discussions (e.g., PubMed Commons[51]), or information sharing portals such as those developed by the ECNP Preclinical Data Forum[52]. In a more general vein, positive findings in one animal strain and negative ones in another should not necessarily be seen as a limitation of a study; rather, they could be used as guidance for much more specifically understanding of the driving force behind the positive findings. Such understanding may lead to the identification of factors which could be used to identify suitable patient populations for clinical testing.

Second, published reports often fail to include important details about the materials and methods used; information that may facilitate the understanding of the origin of discrepant findings. For example, as argued above, mentioning the exact laboratory animal strain nomenclature and the source may be very critical[17]. As it is difficult to establish, *a priori*, what information may turn out to be crucial, the study protocols including all potentially relevant details (e.g., health reports) may eventually be required to be stored in online repositories and referenced in the to-be-published work.

Third, large collaboration projects involving independent laboratories could provide a basis for studying treatment response heterogeneity. A recent EU-funded initiative[53] has established web-based platforms for multicenter animal studies. Another example is the Interventions Testing Program at the National Institute of Aging, which has the explicit aim of confirming the reported potential of treatments to extend lifespan, and delay disease and dysfunction under the most rigorous conditions — multiple test sites, specially bred genetically heterogeneous mice of both sexes, and very well powered[54].

It has to be emphasized that such multi-site research projects often aim at generating more robust findings[55] and, therefore, need to be equipped with additional resources and specific technologies to enable analysis of response heterogeneity. On the one hand, thanks to advances in technologies and reduced costs, whole-genome sequencing is no longer considered a technical problem, even for preclinical studies. However, computational methodologies for data analysis and interpretation of the functional relevance of identified genetic variants may present a challenge requiring careful planning and appropriate funding[56]. For example, in a study conducted in 23 inbred mouse strains, a number of cardiovascular parameters were assessed after chronic administration of a β-adrenoceptor agonist and antagonist. Reflecting the complexity of the observed patterns of effects, the conclusion was that "cardiovascular phenotypes are unlikely to segregate according to global phylogeny, but rather be governed by smaller, local differences in the genetic architecture of the various strains"[57]. On the other hand, in order to reliably identify genomic predictors of drug response or to effectively identify a drug's mechanism of action in a multi-site project, standardization of drug-response measurements is essential[58] and, if insufficient, may lead to discrepant observations that are too weak to support precision medicine.

In conclusion, heterogeneity between models and experimental procedures are an important source of the lack of robustness of the reported research findings. Such unrobust data can be a relevant cause of failure in translating preclinical findings into clinical efficacy, and subsequent attrition and exaggerated cost in drug development. The flipside of this coin is that understanding of the causes of heterogeneity may lead to the identification of relevant factors for the identification of responder populations. Thus, heterogeneity within preclinical findings, if built into study design, controlled for and evaluated effectively, may prove to be an asset, rather than a problem, for the development of precision medicine. To enable this paradigm shift, several steps need to be taken to identify conditions under which drugs do not work. An improved granularity in the reporting of preclinical

studies is central among them (i.e., details about study design, experimental conditions, quality of tools and reagents, validation of assay conditions, etc.). These actions need to be discussed jointly by preclinical data robustness and precision medicine research communities; emphasizing the role of collaboration in success strategies[59,60].

In closing, we would like to draw attention to an example of how careful attention to the differences between research models may lead to important discoveries. The T-cell-derived S49 cell line can be killed by β-adrenoceptor agonists, such as isoprenaline, and other cAMP- increasing agents, such as forskolin. Sub-strains of this cell line have been shown to be differentially sensitive to those agents[61]. Looking more deeply into these findings has led to the identification of G-proteins and adenylyl cyclases as parts of the cellular signal transduction machinery, and eventually to a Nobel Prize for Alfred G. Gilman, the scientist who observed and reported these differences.

Conflict of Interest and Funding

No conflict of interest was reported by all authors.

Acknowledgements

The authors would like to thank Dr. Karen Chu for stimulating discussions and help with editing the style and language of the manuscript.

Reference

1. Scannell J W, Blanckley A, Boldon H, *et al*., 2012, Diagnosing the decline in pharmaceutical R&D efficiency. *Nature Reviews Drug Discovery*, vol.11: 191–200. http://dx.doi.org/10.1038/nrd3681
2. Herper M, 2012, *The truly staggering cost of inventing new drugs*, viewed April 16, 2016, <http://onforb.es/yNffHT>
3. Grainger D, 2015, *Why too many clinical trials fail -- And a simple solution that could increase returns on pharma R&D*, viewed April 16, 2016, <http://onforb.es/15Vtfe7>
4. Kola I and Landis J, 2004, Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, vol.3: 711–716. http://dx.doi.org/10.1038/nrd1470
5. Garner J P, 2014, The significance of meaning: why do over 90% of behavioral neuroscience results fail to translate to humans, and what can we do to fix it? *ILAR Journal*, vol.55(3): 438–456. http://dx.doi.org/10.1093/ilar/ilu047
6. Millan M J, Goodwin G M, Meyer-Lindenberg A, *et al*., 2015, Learning from the past and looking to the future: Emerging perspectives for improving the treatment of psychiatric disorders. *European Neuropsychopharmacology*, vol.25(5): 599–656. http://dx.doi.org/10.1016/j.euroneuro.2015.01.016
7. Köster U, Nolte I and Michel M C, 2016, Preclinical research strategies for newly approved drugs as reflected in early publication patterns. *Naunyn-Schmiedeberg's Archives of Pharmacology*, vol.389(2): 187–199. http://dx.doi.org/10.1007/s00210-015-1187-1
8. Insel T R, Voon V, Nye J S, *et al*., 2013, Innovative solutions to novel drug development in mental health. *Neuroscience and Biobehavioral Reviews*, vol.37(10 Part 1): 2438–2444. http://dx.doi.org/10.1016/j.neubiorev.2013.03.022
9. Tymianski M, 2015, Neuroprotective therapies: preclinical reproducibility is only part of the problem. *Science Translational Medicine*, vol.7(299): 299fs32. http://dx.doi.org/10.1126/scitranslmed.aac9412
10. Williams M, 2011, Productivity shortfalls in drug discovery: contributions from the preclinical sciences? *The Journal of Pharmacology and Experimental Therapeutics*, vol.336(1): 3–8. http://dx.doi.org/10.1124/jpet.110.171751
11. Prinz F, Schlange T and Asadullah K, 2011, Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, vol.10: 712–713. http://dx.doi.org/10.1038/nrd3439-c1
12. Begley C G and Ellis L M, 2012, Drug development: raise standards for preclinical cancer research. *Nature*, vol.483: 533–533. http://dx.doi.org/10.1038/483531a
13. Collins F S and Tabak L A, 2014, Policy: NIH plans to enhance reproducibility. *Nature*, vol.505(7485): 612–613. http://dx.doi.org/10.1038/505612a
14. Freedman L P, Cockburn I M and Simcoe T S, 2015, The economics of reproducibility in preclinical research. *PLoS Biology*, vol.13(6): e1002165. http://dx.doi.org/10.1371/journal.pbio.1002165
15. Colquhoun D, 2014, An investigation of the false discovery rate and the misinterpretation of *p*-values. *Royal Society Open Science*, vol.1: 140216. http://dx.doi.org/10.1098/rsos.140216
16. Motulsky H J, 2014, Common misconceptions about data analysis and statistics. *Naunyn-Schmiedeberg's Archives of Pharmacology*, vol.387(11): 1017–1023. http://dx.doi.org/10.1007/s00210-014-1037-6
17. McGrath J C, Drummond G B, McLachlan E M, *et al*., 2010, Guidelines for reporting experiments involving animals: the ARRIVE guidelines. *British Journal of*

Pharmacology, vol.160(7): 1573–1576.
http://dx.doi.org/10.1111/j.1476-5381.2010.00873.x

18. Curtis M J, Bond R A, Spina D, *et al*., 2015, Experimental design and analysis and their reporting: new guidance for publication in BJP. *British Journal of Pharmacology*, vol.172(14): 3461–3471.
http://dx.doi.org/10.1111/bph.12856
19. Jarvis M F and Williams M, 2016, Irreproducibility in preclinical biomedical research: perceptions, uncertainties, and knowledge gaps. *Trends in Pharmacological Sciences*, vol.37(4): 290–302.
http://dx.doi.org/10.1016/j.tips.2015.12.001
20. Alberts B, Cicerone R J, Fienberg S E, *et al*., 2015, Self-correction in science at work. *Science*, vol.348(6242): 1420–1422.
http://dx.doi.org/10.1126/science.aab3847
21. Clemens M A, 2015, The meaning of failed replications: a review and proposal. *Journal of Economic Surveys*, 1–17.
http://dx.doi.org/10.1111/joes.12139
22. Kenett R S and Shmueli G, 2015, Clarifying the terminology that describes scientific reproducibility. *Nature Methods*, vol.12: 699.
http://dx.doi.org/10.1038/nmeth.3489
23. Llovera G, Hofmann K, Roth S, *et al*., 2015, Results of a preclinical randomized controlled multicenter trial (pRCT): anti-CD49d treatment for acute brain ischemia. *Science Translational Medicine*, vol.7(299): 299ra121.
http://dx.doi.org/10.1126/scitranslmed.aaa9853
24. Dale P R, Cernecka H, Schmidt M, *et al*., 2014, The pharmacological rationale for combining muscarinic receptor antagonists and β-adrenoceptor agonists in the treatment of airway and bladder disease. *Current Opinion in Pharmacology*, vol.16: 31–42.
http://dx.doi.org/10.1016/j.coph.2014.03.003
25. Maman K, Aballea S, Nazir J, *et al*., 2014, Comparative efficacy and safety of medical treatments for the management of overactive bladder: a systematic literature review and mixed treatment comparison. *European Urology*, vol.65(4): 755–765.
http://dx.doi.org/10.1016/j.eururo.2013.11.010
26. Michel M C, Brunner H R, Foster C, *et al*., 2016, Angiotensin II type 1 receptor antagonists in animal models of vascular, cardiac, metabolic and renal disease. *Pharmacology and Therapeutics* (in press).
http://dx.doi.org/10.1016/j.pharmthera.2016.03.019
27. Lindpaintner K, Kreutz R and Ganten D, 1992, Genetic variation in hypertensive and 'control' strains: what are we controlling for anyway? *Hypertension*, vol.19: 428–430.
http://dx.doi.org/10.1161/01.HYP.19.5.428
28. Kumar G, Talpos J and Steckler T, 2015, Strain-dependent effects on acquisition and reversal of visual and spatial tasks in a rat touchscreen battery of cognition. *Physiology and Behavior*, vol.144: 26–36.
http://dx.doi.org/10.1016/j.physbeh.2015.03.001
29. Bottger A, den Bieman M, Lankhorst Æ, *et al*., 1996, Strain-specific response to hypercholesterolaemic diets in the rat. *Laboratory Animals*, vol.30(2): 149–157.
http://dx.doi.org/10.1258/002367796780865736
30. Bouxsein M L, Myers K S, Shultz K L, *et al*., 2005, Ovariectomy-induced bone loss varies among inbred strains of mice. *Journal of Bone and Mineral Research*, vol.20(7): 1085–1092.
http://dx.doi.org/10.1359/JBMR.050307
31. Iwaniec U T, Yuan D, Power R A, *et al*., 2006, Strain-dependent variations in the response of cancellous bone to ovariectomy in mice. *Journal of Bone and Mineral Research*, vol.21(7): 1068–1074.
http://dx.doi.org/10.1359/jbmr.060402
32. Cui Q, Hodgetts S I, Hu Y, *et al*., 2007, Strain-specific differences in the effects of cyclosporin A and FK506 on the survival and regeneration of axotomized retinal ganglion cells in adult rats. *Neuroscience*, vol.146(3): 986–999.
http://dx.doi.org/10.1016/j.neuroscience.2007.02.034
33. Coruzzi G, Pozzoli C, Adami M, *et al*., 2012, Strain-dependent effects of the histamine H_4 receptor antagonist JNJ7777120 in a murine model of acute skin inflammation. *Experimental Dermatology*, vol.21(1): 32–37.
http://dx.doi.org/10.1111/j.1600-0625.2011.01396.x
34. Maronpota R R, Thoolen R J M M and Hansen B, 2015, Two-year carcinogenicity study of acrylamide in Wistar Han rats with in utero exposure. *Experimental and Toxicologic Pathology*, vol.67(2): 189–195.
http://dx.doi.org/10.1016/j.etp.2014.11.009
35. Kawedia J D, Janke L, Funk A J, *et al*., 2012, Substrain-specific differences in survival and osteonecrosis incidence in a mouse model. *Comparative Medicine*, vol.62(6): 466–471.
36. Palm S, Roman E and Nylander I, 2012, Differences in basal and ethanol-induced levels of opioid peptides in Wistar rats from five different suppliers. *Peptides*, vol.36(1): 1–8.
http://dx.doi.org/10.1016/j.peptides.2012.04.016
37. International Committee on Standardized Genetic Nomenclature for Mice and Rat Genome and Nomenclature Committee, *Guidelines for nomenclature of mouse and rat strains*, revised January 2016, viewed April 16, 2016, <http://www.informatics.jax.org/mgihome/nomen/strains.shtml>
38. Lindemann L L, Gatti S, Ballard T, *et al*., 2006, *Neuroscience meeting planner, October 14–18, 2006: Pharmacological and biochemical characterization of a population of Wistar rats with reduced expression of metabo-*

tropic glutamate receptor 2 protein (mGluR2). Society for Neuroscience, Atlanta, GA.

39. Ceolin L, Kantamneni S, Barker G R I, *et al*., 2011, Study of novel selective mGlu2 agonist in the temporo-ammonic input to CA1 neurons reveals reduced mGlu2 receptor expression in a Wistar substrain with an anxiety-like phenotype. *The Journal of Neuroscience*, vol.31(18): 6721–6731. http://dx.doi.org/10.1523/JNEUROSCI.0418-11.2011
40. Wood G K, Marcotte E R, Quirion R, *et al*., 2001, Strain differences in the behavioural outcome of neonatal ventral hippocampal lesions are determined by the postnatal environment and not genetic factors. *European Journal of Neuroscience*, vol.14(6): 1030–1034. http://dx.doi.org/10.1046/j.0953-816x.2001.01716.x
41. Klengel T and Binder E B, 2013, Gene × environment interactions in the prediction of response to antidepressant treatment. *International Journal of Neuropsychopharmacology*, vol.16(3): 701–711. http://dx.doi.org/10.1017/S1461145712001459
42. *Alkermes announces topline results of FORWARD-3 and FORWARD-4, two phase 3 studies of ALKS 5461 in major depressive disorder*, n.d., viewed April 16, 2016, <http://www.businesswire.com/news/home/20160121005348/en/Alkermes-Announces-Topline-Results-FORWARD-3-FORWARD-4-Phase>
43. Carr G V, Bangasser D A, Bethea T, *et al*., 2010, Antidepressant-like effects of κ-opioid receptor antagonists in Wistar Kyoto rats. *Neuropsychopharmacology*, vol.35: 752–763. http://dx.doi.org/10.1038/npp.2009.183
44. Kinon B J, Millen B A, Zhang L, *et al*., 2015, Exploratory analysis for a targeted patient population responsive to the metabotropic glutamate 2/3 receptor agonist pomaglumetad methionil in schizophrenia. *Biological Psychiatry*, vol.78(11): 754–762. http://dx.doi.org/10.1016/j.biopsych.2015.03.016
45. Zhou Z, Karlsson C, Liang T, *et al*., 2013, Loss of metabotropic glutamate receptor 2 escalates alcohol consumption. *Proceedings of the National Academy of Sciences of the United States of America*, vol.110(42): 16963–16968. http://dx.doi.org/10.1073/pnas.1309839110
46. Meinhardt M W, Hansson A C, Perreau-Lenz S, *et al*., 2013, Rescue of infralimbic mGluR2 deficit restores control over drug-seeking behavior in alcohol dependence. *The Journal of Neuroscience*, vol.33(7): 2794–2806. http://dx.doi.org/10.1523/JNEUROSCI.4062-12.2013
47. Michel M C and Korstanje C, 2016, β3-Adrenoceptor agonists for overactive bladder syndrome: role of translational pharmacology in a repositioning drug development project. *Pharmacology and Therapeutics*, vol.159: 66–82. http://dx.doi.org/10.1016/j.pharmthera.2016.01.007
48. Larsen T M, Toubro S, van Baak M A, *et al*., 2002, Effect of a 28-d treatment with L-796568, a novel β3-adrenergic receptor agonist, on energy expenditure and body composition in obese men. *The American Journal of Clinical Nutrition*, vol.76(4): 780–788.
49. *Preclinical Reproducibility and Robustness*, n.d., F1000 Research, viewed June 14, 2016, <http://f1000research.com/channels/PRR>
50. *bioRxiv*, n.d., Cold Spring Harbor Laboratory, viewed June 16, 2016, <http://biorxiv.org>
51. *PubMed Commons*, n.d., viewed June 17, 2016, <http://www.ncbi.nlm.nih.gov/pubmedcommons>
52. *Preclinical Data Forum Network*, n.d., ECNP, viewed June 17, 2016, <https://www.ecnp.eu/projects-initiatives/ECNP-networks/List-ECNP-Networks/Preclinical-Data-Forum.aspx>
53. *Multi-PART (Multicentre Preclinical Animal Research Team)*, n.d., viewed June 8, 2016, <http://www.dcn.ed.ac.uk/multipart>
54. *Interventions Testing Program (ITP)*, n.d., National Institute on Aging, viewed June 11, 2016, <https://www.nia.nih.gov/research/dab/interventions-testing-program-itp>
55. Richter S H, Garner J P, Auer C, *et al*., 2010, Systematic variation improves reproducibility of animal experiments. *Nature Methods*, vol.7: 167–168. http://dx.doi.org/10.1038/nmeth0310-167
56. Lauschke V M and Ingelman-Sundberg M, 2016, Precision medicine and rare genetic variants. *Trends in Pharmacological Science*, vol.37(2): 85–86. http://dx.doi.org/10.1016/j.tips.2015.10.006
57. Berthonneche C, Peter B, Schüpfer F, *et al*., 2009, Cardiovascular response to beta-adrenergic blockade or activation in 23 inbred mouse strains. *PLoS One*, vol.4(8): e6610. http://dx.doi.org/10.1371/journal.pone.0006610
58. Haibe-Kains B, El-Hachem N, Birkbak N J, *et al*., 2013, Inconsistency in large pharmacogenomic studies. *Nature*, vol.504: 389–393. http://dx.doi.org/10.1038/nature12831
59. Landeck L, Lessl M, Reischl J, *et al*., 2016, Collaboration for success: the value of strategic collaborations for precision medicine and biomarker discovery. *Advances in Precision Medicine*, vol.1(1): 25–33. http://dx.doi.org/10.18063/APM.2016.01.002
60. Kannt A and Wieland T, 2016, Managing risks in drug discovery: reproducibility of published findings. *Naunyn-Schmiedeberg's Archives of Pharmacology*, vol.389(4): 353–360. http://dx.doi.org/10.1007/s00210-016-1216-8
61. Gilman A G, 1995, G proteins and regulation of adenylyl cyclase. *Bioscience Reports*, vol.15(2): 65–97.